UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons

Sicheng Yang*
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
yangsc21@mails.tsinghua.edu.cn

Zilin Wang*
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
wangzl21@mails.tsinghua.edu.cn

Zhiyong Wu†
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
The Chinese University of Hong Kong
Hong Kong SAR, China
zywu@sz.tsinghua.edu.cn

Minglei Li†
Huawei Cloud Computing Technologies Co., Ltd
Shenzhen, China
liminglei29@huawei.com

Zhensong Zhang
Huawei Noah's Ark Lab
Shenzhen, China
zhangzhensong@huawei.com

Qiaochu Huang
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
hqc22@mails.tsinghua.edu.cn

Lei Hao
Huawei Noah's Ark Lab
Shenzhen, China
haolei5@huawei.com

Songcen Xu
Huawei Noah's Ark Lab
Shenzhen, China
xusongcen@huawei.com

Xiaofei Wu
Huawei Noah's Ark Lab
Shenzhen, China
wuxiaofei2@huawei.com

Changpeng Yang
Huawei Cloud Computing Technologies Co., Ltd
Shenzhen, China
yangchangpeng@huawei.com

Zonghong Dai
Huawei Cloud Computing Technologies Co., Ltd
Shenzhen, China
daizonghong@huawei.com

ABSTRACT

The automatic co-speech gesture generation draws much attention in computer animation. Previous works designed network structures on individual datasets, which resulted in a lack of data volume and generalizability across different motion capture standards. In addition, it is a challenging task due to the weak correlation between speech and gestures. To address these problems, we present UnifiedGesture, a novel diffusion model-based speech-driven gesture synthesis approach, trained on multiple gesture datasets with different skeletons. Specifically, we first present a retargeting network to learn latent homeomorphic graphs for different motion capture standards, unifying the representations of various gestures while extending the dataset. We then capture the correlation between speech and gestures based on a diffusion model architecture using cross-local attention and self-attention to generate better speech-matched and realistic gestures. To further align speech and gesture and increase diversity, we incorporate reinforcement learning on the discrete gesture units with a learned reward function. Extensive experiments show that UnifiedGesture outperforms recent approaches on speech-driven gesture generation in terms of CCA, FGD, and human-likeness. All code, pre-trained models, databases, and demos are available to the public at https://github.com/YoungSeng/UnifiedGesture.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Motion processing**; **Neural networks**.

KEYWORDS

gesture generation, neural motion processing, data-driven animation

*Equal contribution
†Corresponding author

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612503

ACM Reference Format:
Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Qiaochu Huang, Lei Hao, Songcen Xu, Xiaofei Wu, Changpeng Yang, and Zonghong Dai. 2023. UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3581783.3612503

Figure 1: Gesture examples generated by our proposed method. Different skeletons are unified to the primal skeleton. The speech-driven primal skeleton generate gestures for the specified skeleton. The character used in the paper is publicly available.

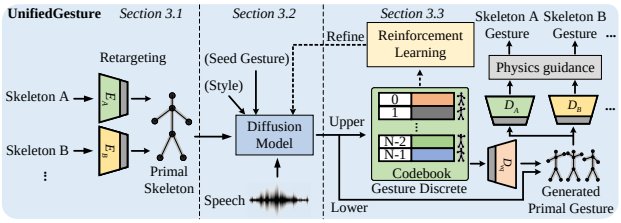

Figure 2: Gesture generation pipeline of our proposed framework. We retarget the different skeletons to the primal skeleton. Given a speech segment (optional style, seed gesture), the output is the primal gesture after VQVAE encoding of the output of the diffusion model. We introduce reinforcement learning to refine the gesture generation network. Finally, a gesture of the skeleton is specified and generated, with physics guidance.

1 INTRODUCTION

Nonverbal behaviors, including gestures, play key roles in conveying messages in human communication [37]. The automatic co-speech gesture generation is considered an enabling technology to create realistic 3D avatars in films, games, virtual social spaces, and for interaction with social robots [56]. In the era of deep learning, existing data-driven gesture generation methods usually rely on a large dataset. Studies have shown that a larger amount of data can improve the generalization of the model and enhance its performance [10, 57].

Thanks to the development of human pose estimation [46], it's easy to extract 3D human poses from tremendous 2D gesture data on the web, e.g., TED [89] and PATS [2], some works [47, 88, 89] are based on 2D gesture datasets. While large in quantity, 3D poses extracted from 2D datasets are poor in quality and difficult to use, most works [4, 17, 36, 43, 49] opt for high quality 3D mocap datasets.

There are two main challenges when utilizing 3D datasets. First, due to the expensive cost of motion capture, the typical 3D gesture datasets [16, 19, 38, 49] are relatively small, thus the generalization of the models trained on the individual dataset is limited, and the ability of the trained algorithms is also confined to the content of the individual dataset. For example, some datasets contain style information [19, 49], while the others do not [16, 38]. Second, it is not straightforward to train algorithms on mixture datasets directly, since different datasets usually have different skeletons, they are captured with different mocap systems. Most of the current solutions use software such as Blender [18] or Maya [8] for automatic retargeting to a unified skeleton, which requires manual specification of the bone mapping and leads to unavoidable errors [1]. The irregular connectivity and hierarchical structure of the skeleton joint motion cause difficulties in the large-scale application of multiple skeletons.

To tackle these challenges, we propose UnifiedGesture, a novel unified co-speech gesture synthesis model for multiple skeletons. The overview of our method is shown in Figure 2. Although the number and position of the different skeleton joints are different, they all correspond to homeomorphic (topologically equivalent) graphs [81]. Unlike sign language or hand gestures, there is a weak correlation between speech and body gestures at a coarse-grained level [61, 84]. Specifically, we assume that the gesture details associated with speech are contained in the primal skeleton gesture. According to this assumption, we first use a data-driven deep skeleton-aware [1] framework to learn latent homeomorphic graphs for different skeletons. The different skeletons are unified and retargeted to the primal skeleton while extending the dataset. Then we introduce a denoising-diffusion-based speech-driven co-speech gesture generation model, using WavLM features [12], based on cross-local attention [64] and self-attention [75] architecture to better capture the temporal information between audio and gestures. Third, unlike speaking with the face or lips, the weak correlation between speech and gesture lacks a suitable criterion for learning the model, to refine the gesture generation model, we employ inverse reinforcement learning (IRL) on discrete gesture units to train a reward model that evaluates the generated gestures and guides the diffusion model to generate high-quality and diverse gestures aligned with speech during the reinforcement learning (RL) process. Our code, pre-trained models, and demos will be publicly available soon. The main contributions of our work are:

- We employ a skeleton-aware retargeting network to unify the different skeletons to a common primal skeleton while extending the dataset.
- We present a temporally aware attention-based diffusion model on the primal skeleton for speech-driven co-speech gesture generation. By virtue of the diffusion model, we can edit the style of the gestures, setting the initial gestures, and generating diverse gestures.
- We introduce reinforcement learning with a learned reward function to refine the generation model and make the model explore the data. The exploratory space for reinforcement learning is reduced by learning a codebook with VQVAE to summarize meaningful gesture units.
- Extensive experiments show that our model can generate human-like, speech-matched, stylized, diverse, controllable, and physically plausible gestures that significantly outperform existing gesture generation methods.

2 RELATED WORK

2.1 Motion Retargeting

Our task is to take advantage of multiple gesture datasets. There are two main challenges. First, different datasets have the same

motion capture standards (e.g., Trinity [16] and BEAT [49] both using Vicon's suits); second, different datasets have different motion capture standards (e.g., ZEGGS [19] and Talking With Hands [41]). For the first case, we can select the body joints common to both datasets and unify the different skeletons normalized [88] by height or arm span, etc. The latter case is more challenging. Ma et al. [54] try to map multiple datasets to a defined skeleton, but it still partially relies on handcrafting and the results are still limited to a specific motion skeleton. Some work [31, 35] try to retarget the motion of different skeletons by VAE, using standard convolution and pooling. However, unlike images or videos, different skeletons exhibit irregular connectivity. Villegas et al. [77] propose a neural network for motion retargeting that adapts input motion to target characters, achieving state-of-the-art results and using cycle consistency for unsupervised learning. Lim et al. [48] propose a pose-movement network for motion retargeting using a normalizing process and novel loss function. Kim et al. [30] present an unsupervised motion retargeting model using temporal dilated convolutions that generates realistic and stable trajectories for humanoid characters. Villegas et al. [76] propose a motion retargeting method that preserves self-contacts and prevents interpenetration, using a recurrent network. Li et al. [44] propose an iterative motion retargeting method using an iterative motion retargeting network for unsupervised motion retargeting. Inspired by [1], we use a deep skeleton-aware framework for data-driven motion retargeting between skeletons.

2.2 Gesture Generation

2.2.1 End-to-end Co-speech Gesture Generation. Gesture generation is a complex task that requires understanding speech, gestures, and their relationships. The present data-driven studies mainly consider four modalities: text [6, 80, 89], audio [20, 24, 62], gesture motion [52, 85, 88], and speaker identity [3, 4, 50]. There are some works to extend the scale of the dataset. Liu et al. [49] present a large motion capture dataset for studying the correlation of conversational gestures with facial expressions, emotions, and semantics. Kucherenko et al. [38] provide further annotation of the specific properties and details of the gestures in the dataset. Ghorbani et al. [19] propose a dataset containing motion styles compared to the previous dataset containing speech styles. However, these methods are currently only tried on a single dataset, which resulted in a lack of data volume and generalizability across different motion capture standards.

Some works [23, 92] use motion-matching methods to generate co-speech gestures. Yang et al. [86] try to transform the original gesture motion space to the deep latent phase space [66], but it is still based on the traditional convolutional network, ignoring the hierarchy and connectivity between the skeletons. Besides, this approach requires careful design of the database, which is directly related to the performance of the generated gestures. The length of matching needs to be balanced between quality and diversity. Furthermore, the approach also requires the complex and time-consuming manual design of the matching rules.

2.2.2 Diffusion Models for Motion Generation. Diffusion models [25] excel at modeling complicated data distribution and generating vivid motion sequences. Many works [29, 63, 72] integrate diffusion-based generative models into the motion domain. There are some works [5, 91, 94] that introduce diffusion models in gesture generation to demonstrate the potential of diffusion models in solving cross-modal, time-series relations problems. In our work, we use a well-designed attention architecture in the diffusion model to make the generated gestures match better with the speech.

2.2.3 Quantization-based Pose Representation. Kipp has represented gestures as predefined unit gestures [33]. Lucas et al. [53] propose to train a GPT-like model for next-index prediction. Li et al. [45] propose to pose VQ-VAE [74] to encode and summarize dancing units. In terms of gesture generation, there are several works [7, 84, 93] that apply VQVAE to encode meaningful gesture units. Existing studies [21, 27, 53] have shown that quantification helps to reduce motion freezing during motion generation and retains the details of motion well. Unlike them, we encode the gesture units in the deep primal motion latent space.

2.3 Reinforcement Learning

The goal of reinforcement learning (RL) is to learn a policy that maximizes rewards through iterative interactions with the environment [69]. The agent takes actions based on the current state, receives rewards, and updates its policy. The trial-and-error learning nature of RL enables it to be a versatile method for making decisions in complex and dynamic environments [15, 42, 78]. RL algorithms can be broadly categorized into two types: value-based [13, 68, 82] and policy-based [9, 22, 83]. Value-based methods estimate expected rewards for actions in a given state. Policy-based algorithms directly learn a policy model that maps states to actions and updates using Policy Gradient [70]. Policy-based RL are popular for handling high-dimensional state and action spaces, and non-differentiable reward functions. Since the rewards in RL do not need to be differentiable with respect to model parameters, RL algorithms can be applied to a wide range of reward maximization problems [45, 58, 60]. Related to our work, Sun et al. [67] propose a contrastive pre-trained reward to evaluate the correspondence between gesture and speech sequences and employs conservative Q-Learning (CQL) [40] for model optimization. Although state transitions and reward functions are obtainable, the method relies on offline RL, which limits the capability of the model. Bailando [45] use a hand-designed reward function to fine-tune the dance generation model. However, hand-designed reward functions require significant expert knowledge and make it difficult to comprehensively evaluate actions. In our work, we fine-tune our diffusion model with online RL on the training set, using the learned reward model to refine the gesture generation model.

3 OUR APPROACH

3.1 Multiple Skeletons Retargeting Network

The structure of a skeleton is typically hierarchical [1], so we use graphs [79] to represent motions. Connectivity is determined by the kinematic chains (the paths from the root joint to the end-effectors). Nodes to represent corresponding joints and, in particular, leaf nodes to represent end-effectors. The adjacency lists are expressed as $\mathcal{N}^d = \left\{\mathcal{N}_1^d, \mathcal{N}_2^d, \ldots, \mathcal{N}_J^d\right\}$, where J is the number of joints and

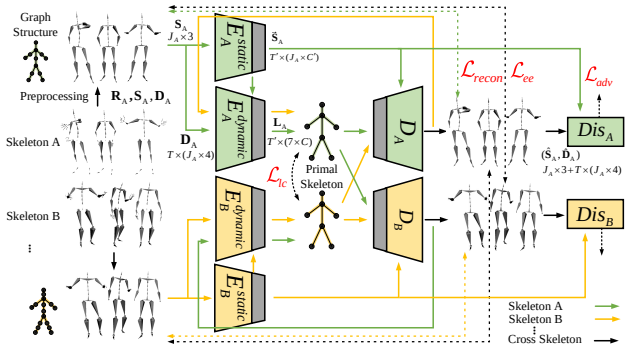

Figure 3: Motion is represented using a static encoder E^{static} and a dynamic encoder $E^{dynamic}$. Assuming that all skeletons contain 5 end-effectors (2 hands, 2 feet, and head) and 2 mid-nodes, we use the latent representation L of the primal skeleton to represent all the different skeletons.

$\mathcal{N}_i^d$ denotes the edges whose distance in the tree is equal or less than d from the i-th edge.

3.1.1 Reference Pose Unification. The current full-body motion capture dataset for speech-driven gestures contains mainly: Trinity [16] (244 min of audio, a male actor), ZEGGS [19] (135 min of audio, a female actor, 19 different motion styles), BEAT [49] (76 hours, 30 speakers, 8 different emotions) and Talking with Hands [41] (50 hours, two-person face-to-face conversations). More details on the different skeletons can be found in the supplementary material. Different gesture datasets have different reference poses and motion representations (number and position of joints). We take two datasets A and B as examples. We first need to set the position and rotation of the unified reference poses, such as T-pose or A-pose. We first centralize the reference representation to the root joints (e.g. the Talking with Hands dataset uses the 'world' joint to maintain height). Two reference poses $\mathbf{P}_A$ and $\mathbf{P}_B$ can be aligned through global and local translation and rotation:

$$\mathbf{P}_B = \mathbf{Q}^{AB}\mathbf{P}_A \left(\mathbf{Q}^{AB}\right)^{\top} \tag{1}$$

where $\mathbf{Q}^{AB}$ denotes the reference poses transfer matrix.

The reference representation of a motion sequence of length T based on reference pose $\mathbf{P}$ can be represented by 3D position and 4D rotation of the root joint as $\mathbf{R} \in \mathbb{R}^{T\times(3+4)}$. The reference representations $\mathbf{R}$ of different skeletons can be retargeted after normalization according to the height of the reference pose.

3.1.2 Motion Unification. The motion of different skeletons consists of a static component $\mathbf{S} \in \mathbb{R}^{J\times 3}$ (joint offsets) and a dynamic one $\mathbf{D} \in \mathbb{R}^{T\times(J\times 4)}$ (joint rotations). To unify the motion of the different skeletons, we utilize a retargeting network architecture similar to [95]. The architecture is shown in Figure 3. Here we take static component $\mathbf{S}_A$ and dynamic component $\mathbf{D}_A$ of skeleton A with J_A joints as an example. First, we adopt skeletal convolution and pooling layers [1] in encoders to extract a deep latent representation $\mathbf{L_A}$ of the motion $\mathbf{S}$ and $\mathbf{D}$, which can be formulated as

$$\overrightarrow{\mathbf{S_A}} = E_A^{static}(\text{repeat } [\mathbf{S}_A] \times T') \tag{2}$$

$$\mathbf{L_A} = E_A^{dynamic}(\mathbf{D}_A, \overrightarrow{\mathbf{S_A}}) \tag{3}$$

where operator 'repeat' denotes tiled and concatenated along the time dimension, $\overrightarrow{\mathbf{S_A}} \in \mathbb{R}^{T'\times(J_A\times C')}$ and $\mathbf{L_A} \in \mathbb{R}^{T'\times(7\times C)}$, $T' = T/d_{re}$, d_{re} is the temporal down-sampling rate. Assuming that all skeletons contain 5 end-effectors (2 hands, 2 feet, and head) and 2 mid-nodes, we use the latent representation $\mathbf{L_A}$ of the primal skeleton to represent all the different skeletons, which contains only 7 nodes. C' and C are the numbers of deep static and dynamic latent channels.

A following skeletal de-convolutional decoder D_A projects $\overrightarrow{\mathbf{S_A}}$ and $\mathbf{L_A}$ back to the motion space as $\hat{\mathbf{S}}_\mathbf{A}$, which can be formulated as

$$(\hat{\mathbf{S}}_\mathrm{A}, \hat{\mathbf{D}}_{\mathrm{A}\to\mathrm{A}}) = D_A(\mathbf{L_A}, \overrightarrow{\mathbf{S_A}}) \tag{4}$$

where $\hat{\mathbf{D}}_{\mathrm{A}\to\mathrm{A}}$ indicates that $\mathbf{L_A}$ is fed into the decoder D_A, simplified as $\hat{\mathbf{D}}_\mathrm{A}$.

During training, D_A tries to reconstruct the input motion, so the decoders are trained by minimizing the reconstruction losses:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}\left[\left\|\hat{\mathbf{D}}_\mathrm{A} - \mathbf{D}_\mathrm{A}\right\|^2\right] + \mathbb{E}\left[\left\|\mathrm{FK}\left(\hat{\mathbf{D}}_\mathrm{A}, \hat{\mathbf{S}}_A\right) - \mathrm{FK}(\mathbf{D}_\mathrm{A}, \mathbf{S}_A)\right\|^2\right] \tag{5}$$

where operator 'FK' is the forward kinematic to get the joint positions, which prevents the accumulaiton of error along the kinematic chain [59].

The skeletal-aware encodes enables retargeting motions of different skeletons into a common deep primal skeleton latent space. A latent consistency loss is applied to this shared representation to ensure that the retargeted motion retains the same dynamic features as the original clip:

$$\mathcal{L}_{\text{lc}} = \mathbb{E}\left[\left\|E_B^{dynamic}(\hat{\mathbf{D}}_{\mathrm{A}\to\mathrm{B}}, \overrightarrow{\mathbf{S_B}}) - \mathbf{L_A}\right\|_1\right], \tag{6}$$

where $\hat{\mathbf{D}}_{\mathrm{A}\to\mathrm{B}}$ indicates that $\mathbf{L_A}$ is fed into the decoder D_B.

Since different skeletons can share the same set of end-effectors (typically head, left hand, right hand, left foot, and right foot), the end-effectors of the original skeleton and the retargeted skeleton should have the same normalized velocity to avoid the artifact of re-targeting, such as foot sliding. This can be formulated as

$$\mathcal{L}_{\text{ee}} = \mathbb{E}\sum_{e\in\mathcal{E}}\left\|\frac{V_{A_e}}{h_A} - \frac{V_{B_e}}{h_B}\right\|^2 \tag{7}$$

where V_{A_e} and V_{B_e} are the velocities of the e-th end-effector of skeletons A and B, respectively. $\mathcal{E}$ is the set of end-effectors. h_A and h_B are the height of skeletons A and B, respectively.

And we use discriminator Dis_B to evaluate whether the retargeted motion is plausible. The adversarial loss can be formulated as

$$\mathcal{L}_{\text{adv}} = \mathbb{E}\left[\left\|Dis_B\left(\hat{\mathbf{D}}_{A\to B}, \overrightarrow{\mathbf{S_B}}\right)\right\|^2\right] + \mathbb{E}\left[\left\|1 - Dis_B\left(\hat{\mathbf{D}}_\mathrm{B}, \overrightarrow{\mathbf{S_B}}\right)\right\|^2\right] \tag{8}$$

The loss of the retargeting network can be computed as:

$$\mathcal{L}_{re} = \mathcal{L}_{\text{rec}} + \lambda_{\text{lc}}\mathcal{L}_{\text{lc}} + \lambda_{\text{ee}}\mathcal{L}_{\text{ee}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \tag{9}$$

For details on the network structure, please refer to our supplementary material.

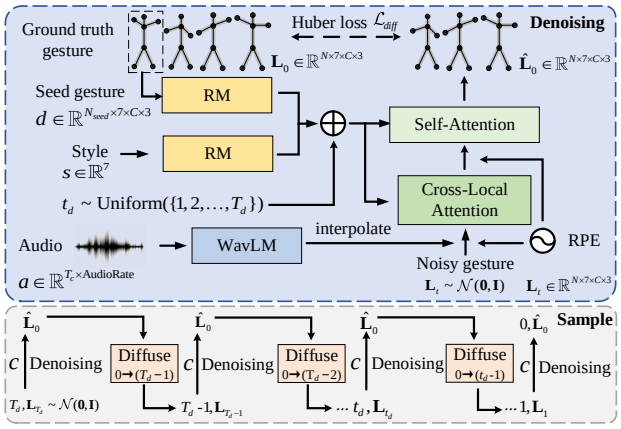

Figure 4: (Top) Denoising module. A noising step t_d and a noisy gesture sequence $\mathbf{L}_t$ at this noising step conditioning on c (including seed gesture d, style s and audio a) are fed into the model. 'RM' is short for random mask. (Bottom) Sample module. At each step t_d, we predict the $\hat{\mathbf{L}}_0$ with the denoising process based on the corresponding conditions, then add the noise to the noising step $\mathbf{L}_{t_d-1}$ with the diffuse process. This process is repeated from $t_d = T_d$ until $t_d = 0$.

3.2 Diffusion Model for Speech-driven Gesture Generation

Diffusion models [25] have made great progress in motion generation [72] due to their ability of to learn to gradually denoising starting from pure noise. We unified the gestures by retargeting the skeletons of different gesture datasets to a primal skeleton, and now obtained a multi-deep primal skeleton gesture set $[\mathbf{L_A}, \mathbf{L_B}, ...]$ with the corresponding speech set $[\mathbf{A_A}, \mathbf{A_B}, ...]$. To generate co-speech gestures with a diffusion model, we use DiffuseStyleGesture [85], which has recently achieved strong results on a single dataset, as our backbone model. As shown in Figure 4, the diffusion model consists of two parts: the forward process (diffusion process) q and the reverse process (denoising process) p_θ.

We denote the generated gesture as $\mathbf{L}$ in the diffusion process, which has the same dimension as an observation data $\mathbf{L}_0 \sim q(\mathbf{L}_0)$, $q(\mathbf{L}_0)$ denotes the distribution of the real data $[\mathbf{L_A}, \mathbf{L_B}, ...]$. According to a variance schedule $\beta_1, \beta_2, \ldots, \beta_{T_d}$ ($0 < \beta_1 < \beta_2 < \cdots < \beta_{T_d} < 1$, T_d is the total time step), we add Gaussian noise

$$q\left(\mathbf{L}_{t_d} \mid \mathbf{L}_{t_d-1}\right) = \mathcal{N}\left(\mathbf{L}_{t_d}; \sqrt{1-\beta_{t_d}}\mathbf{L}_{t_d-1}, \beta_{t_d}\mathbf{I}\right) \tag{10}$$

In denoising process, the denoising process p_θ is a process of learning parameter θ via a neural network. The noise $\mathbf{L}_{t_d}$ at time t_d is used to learn μ_θ, Σ_θ, then

$$p_\theta\left(\mathbf{L}_{t_d-1} \mid \mathbf{L}_{t_d}\right) = \mathcal{N}\left(\mathbf{L}_{t_d-1}; \mu_\theta\left(\mathbf{L}_{t_d}, t_d\right), \Sigma_\theta\left(\mathbf{L}_{t_d}, t_d\right)\right) \tag{11}$$

3.2.1 Denoising Module. Our goal is to synthesize a gesture $\mathbf{L}^{1:N}$ of length N given noising step t_d, noisy gesture $\mathbf{L}_{t_d}$ and conditions c (including audio a, style s, and seed gesture d).

$$\hat{\mathbf{L}}_0 = \text{Denoise}\left(\mathbf{L}_{t_d}, t_d, c\right) \tag{12}$$

During training, noising step t_d is sampled from a uniform distribution of $\{1, 2, \ldots, T_d\}$, with the same position encoding as [75]. Noisy gesture $\mathbf{L}_{t_d}$ has the same dimension as the real gesture $\mathbf{L}_0$ obtained by sampling from the standard normal distribution $\mathcal{N}(0, \mathbf{I})$. In the latent representation of the gesture we also extract the difference between two frames as latent velocity and also extract the difference between two frames of latent velocity as latent acceleration, therefore $\mathbf{L}_0 \in \mathbb{R}^{N\times(7\times C\times3)}$. Audio features are generated from the pre-trained models of WavLM Large [12]. Then we use linear interpolation to align WavLM features and gesture $\mathbf{L}_0$ in the time dimension. The styles of gestures are represented as one-hot vectors where only one element of a selected style is nonzero. Seed gesture helps to make smooth transitions between consecutive syntheses [88]. The first N_{seed} frames of the gestures clip are used as the seed gesture d and the remaining N frames are used as the real gesture $\mathbf{L}_0$ to calculate loss. Self-attention [75] and cross-local attention [64] based on relative position encoding (RPE) [34] are used to generate better speech-matched and realistic gesture. Random masks (RM) are added to the pipeline of seed gesture d and style s feature processing for classifier-free learning [26]. During the training process, we combine the predictions of the conditional model Denoise $\left(\mathbf{L}_{t_d}, t_d, c_1\right), c_1 = [d, s, a]$ and the unconditional model Denoise $\left(\mathbf{L}_{t_d}, t_d, c_2\right), c_2 = [\varnothing, \varnothing, a]$:

$$\hat{\mathbf{L}}_{0\gamma,c_1,c_2} = \gamma\,\text{Denoise}\left(\mathbf{L}_{t_d}, t_d, c_1\right) + (1-\gamma)\,\text{Denoise}\left(\mathbf{L}_{t_d}, t_d, c_2\right) \tag{13}$$

Then, as for style s in condition, we can generate style-controlled gestures when sampling by interpolating or even extrapolating the two variants using γ, as $c_1 = [d, s_1, a], c_2 = [d, s_2, a]$ in Equation (13).

The Denoising module can be trained by optimizing the Huber loss [28] between the generated poses $\hat{\mathbf{L}}_0$ and the ground truth human gestures $\mathbf{L}_0$ on the training examples:

$$\mathcal{L}_{diff} = \lambda_{diff} E_{\mathbf{L}_0\sim q(\mathbf{L}_0|c), t_d\sim[1,T_d]}\left[\text{HuberLoss}(\mathbf{L}_0 - \hat{\mathbf{L}}_0)\right] \tag{14}$$

3.2.2 Sample Module. The final co-speech gesture is given by splicing a number of clips of time duration T_c with frame length N. The initial noisy gesture $\mathbf{L}_{T_d}$ is sampled from the standard normal distribution and the other $\mathbf{L}_{t_d}(t_d < T_d)$ is the result of the previous noising step. The seed gesture for the first clip can be generated by randomly sampling a gesture from the dataset or by setting it to the average gesture. Then the seed gesture for other clips is the last N_{seed} frames of the gesture generated in the previous clip. For every clip, in every noising step t, we predict the clean gesture $\hat{\mathbf{L}}_0$ =Denoise$(\mathbf{L}_{t_d}, t_d, c)$, and add the noise to the noising step $\mathbf{L}_{t_d-1}$ using Equation (10) with the diffuse process. This process is repeated from $t_d = T_d$ until $\mathbf{L}_0$ is reached (Figure 4 bottom). Please refer to our supplementary material for training details such as network structure and implementation details.

3.3 Gesture Generation Refinement

3.3.1 Primal Gesture VQVAE. Here we train a VQVAE to summarize meaningful gesture units to reduce the exploration space for following reinforcement learning. Each code represents a unique gesture. Besides, discrete spaces are more conducive to reinforcement learning for exploration [14, 71]. The architecture of the primal gesture VQVAE is shown in Figure 5. Given the primal gesture

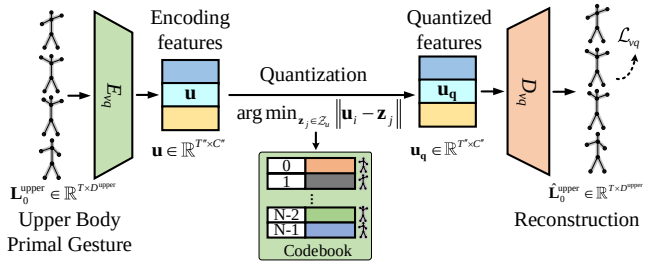

Figure 5: Structure of primal gesture VQVAE. After learning the discrete latent representation of the primal gesture of upper body, the gesture VQVAE encode and summarize meaningful gesture units.

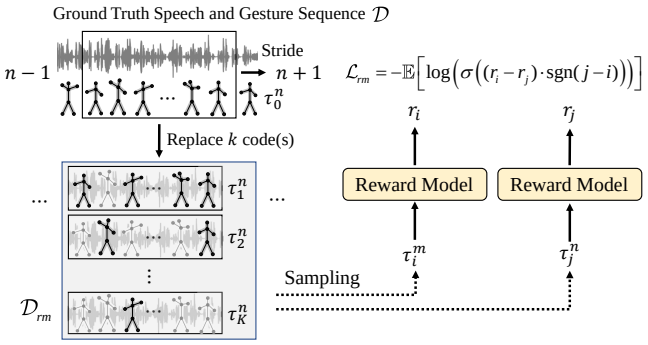

Figure 6: Reward model training. We first sample a VQVAE-encoded speech-gesture pair, denoted as trajectory τ_0. Then, we randomly replace k gesture code(s) with random codes, where $k = 1, \cdots, K$, resulting in K speech-gesture trajectories with decreasing quality. Finally, we utilize the output of reward model r to classify the trajectories with different qualities and optimize the reward model with the loss function $\mathcal{L}_{rm}$.

sequence $\mathbf{L}_0^{\text{upper}} \in \mathbb{R}^{T \times D^{\text{upper}}}$ of the upper body, where D^{upper} denotes primal gesture dimension of the upper body. We first adopt a 1D temporal convolution network E_{vq} to encode the sequence $\mathbf{L}_0^{\text{upper}}$ to context-aware features $\mathbf{u}$

$$\mathbf{u} = E_{vq}(\mathbf{L}_0^{\text{upper}}) \tag{15}$$

where $\mathbf{u} \in \mathbb{R}^{T'' \times C''}$ and $T'' = T/d_{vq}$, d_{vq} is the temporal downsampling rate in VQVAE and C'' is the channel dimension of features. Then we quantize $\mathbf{u}$ by mapping each temporal feature $\mathbf{u}_i$ to its closest codebook [74] element z_j as $\mathbf{q}(.)$:

$$\mathbf{u}_{\mathbf{q},i} = \mathbf{q}(\mathbf{u}) = \arg\min_{\mathbf{z}_j \in \mathcal{Z}_u} \left\| \mathbf{u}_i - \mathbf{z}_j \right\| \tag{16}$$

where $\mathcal{Z}_u$ is a set of C_b codes of dimension n_z. And $\mathbf{u_q}$ is the elements of codebook $\mathcal{Z}_u$, $\mathbf{u}_q \in \mathcal{Z}_u$. A following de-convolutional decoder D_{vq} projects $\mathbf{u_q}$ back to the deep latent space as a primal gesture sequence $\hat{\mathbf{L}}_0^{\text{upper}}$ for the upper body, which can be formulated as

$$\hat{\mathbf{L}}_0^{\text{upper}} = D_{vq}\left(\mathbf{u}_q\right) \tag{17}$$

The VQVAE can be trained by optimizing $\mathcal{L}_{vq}$:

$$\begin{aligned}\mathcal{L}_{vq} = &\|\hat{\mathbf{L}}_0^{\text{upper}} - \mathbf{L}_0^{\text{upper}}\|_1 + \alpha_1 \left\|\hat{\mathbf{L}}_0^{\text{upper}\prime} - \mathbf{L}_0^{\text{upper}\prime}\right\|_1 \\ &+ \alpha_2 \left\|\hat{\mathbf{L}}_0^{\text{upper}\prime\prime} - \mathbf{L}_0^{\text{upper}\prime\prime}\right\|_1 + \left\|\text{sg}[\mathbf{u}] - \mathbf{u}_\mathbf{q}\right\| + \beta_{vq} \left\|\mathbf{u} - \text{sg}\left[\mathbf{u_q}\right]\right\|\end{aligned} \tag{18}$$

where the first item is the reconstruction loss. The next two items are velocity loss and acceleration loss [45, 86]. sg$[\cdot]$ denotes the stop-gradient operation, and the term $\left\|\mathbf{u} - \text{sg}\left[\mathbf{u_q}\right]\right\|$ is the "commitment loss [74]" with weighting factor β_{vq}.

3.3.2 Reinforcement Learning Finetuning. To further enhance the alignment between the speech and gesture and increase the diversity of the generated gestures, we employed reinforcement learning to fine-tune the gesture generation model. The reward signal is pivotal in balancing exploration and exploitation in reinforcement learning. Previous work [45] attempts to optimize partial performance metrics of the model through hand-designed reward functions. However, in our experience, designing heuristic reward functions that comprehensively evaluate the model's performance is challenging. Reinforcement learning training is less stable than supervised learning, and if the reward function only considers specific metrics while neglecting others, the model's overall performance may deteriorate.

In this paper, we adopted Inverse Reinforcement Learning (IRL) [55] to learn a neural network model from human demonstrations to fit the true reward function and explain human behavior. Specifically, our reward model training is shown in Figure 6, similar to [11]. Firstly, we sample a speech-gesture pair from the VQVAE-encoded dataset $\mathcal{D}$, denoted as trajectory τ_0. Then we randomly replace k codes in the trajectory τ_0 where $k = 1, \cdots, K$ to generate K trajectories $[\tau_1, \cdots, \tau_K]$. We sample L tuples and thus get $L \times K$ trajectories to form the dataset $\mathcal{D}_{rm}$ to train the reward model. We make a weak assumption that the more codes replaced with random codes, the worse the quality of the trajectories, including alignment with speech and diversity. Then, we let the reward model R_ψ classify these trajectories with different qualities (may come from different human demonstrations with different speech) $r = R_\psi(\tau)$ to determine which trajectory is better:

$$\mathcal{L}_{rm} = -\mathbb{E}\left[\log\left(\sigma\left((r_i - r_j) \cdot \text{sgn}(j - i)\right)\right)\right], \tag{19}$$

where $\{i, j \in [1, \cdots, K], i \neq j\}$, σ means the sigmoid function and sgn means the signum function:

$$\text{sgn}(x) = \begin{cases} -1, & x < 0 \\ 1, & x > 0 \end{cases}. \tag{20}$$

By learning the classification task, the reward model can learn to output a scalar reward signal $r(\tau) = R_\psi(\tau)$ that makes reasonable evaluations on the quality of the trajectory τ.

Given the reward model, we use the REINFORCE algorithm [68] to improve the model:

$$\mathcal{L}_{RL} = -\mathbb{E}_{\tau \sim \pi}\left[\log p_\pi(\tau) r(\tau)\right], \tag{21}$$

where π means the current policy, *i.e.*, the gesture model and $p_\pi(\tau)$ means the probability of τ given policy π. During the fine-tuning process of the model, the reward model accurately scores the gesture

under the given speech to improve alignment between speech and increase gesture diversity.

3.3.3 Physics Guidance. Inspired by [73], we consider that the foot should have contact with the ground when there is a left-right acceleration or an upward acceleration of the root. Then we use standard Inverse Kinematics (IK) optimization for physics guidance. For more details please refer to the supplementary material.

4 EXPERIMENTS

4.1 Experiment Preparation

4.1.1 Implementation Details. We perform the training and evaluation on the Trinity [16] and ZEGGS [19] datasets. Even based on motion capture, the hand quality is still low [5, 56, 90], so we ignore hand motion currently. Then the number of joints for the two datasets is $J_A = 26$ and $J_B = 27$, respectively. We choose seven more typical and longest-duration styles (happy, sad, neutral, old, relaxed, angry, still) for training and validation. For the Trinity dataset, there are no style labels and we consider all of its styles to be 'neutral'. And we divided the data into 8:1:1 by training, validation, and testing. We first resample the motion of both datasets to 30fps. All audio recordings are downsampled to 16kHz. In terms of retargeting network, we set $d_{re} = 4$, then the primal gesture is 7.5 fps. We set all reference poses $\mathbf{R}$ to the T-pose at the origin with the foot in the Z-plane. The dimension C of each node of the primal gesture in latent space after convolution is 16. We set $\lambda_{\text{lc}} = 1$, $\lambda_{\text{ee}} = 2$ and $\lambda_{\text{adv}} = 0.25$ for Equation (9) and use the Adam [32] optimizer with a batch size of 256 for 16000 epochs. The retargeting network trained on an NVIDIA V100 GPU takes about 3 days. While training the diffusion model and VQVAE, gesture data are cropped to a length of $N = 30$ (4 seconds). For the diffusion model, the Denoising module learns both the conditioned and the unconditioned distributions by randomly masking 10% of the samples using Bernoulli masks. The cross-local attention networks use 8 heads, 32 attention channels, 256 channels, the window size is 6, each window looks at the one window before it, and with a dropout of 0.1. As for self-attention networks are composed of 8 layers, 8 heads, 32 attention channels, 256 channels, and with a dropout of 0.1. We use the AdamW [51] optimizer (learning rate is 3×10^{-5}) with a batch size of 256 for 1000000 steps. Our models have been trained with $T = 1000$ noising steps and a cosine noise schedule. The diffusion model can be learned in about 3 days on one NVIDIA V100 GPU. As for VQVAE, the size C_b of codebook $\mathcal{Z}_u$ is set to 512 with dimension n_z is 512. We set the down-sampling rate $d_{vq} = 2$. And β_{vq}= 0.1, α_1= 1 and α_2= 1 for Equation (18). we use the ADAM optimizer (learning rate is e-4, $\beta_1 = 0.5$, $\beta_2 = 0.98$) with a batch size of 128 for 200 epochs. The VQVAE is learned on one NVIDIA A100 GPU for several hours. For more datasets and training details please refer to the supplementary material.

4.1.2 Evaluation Metrics. Canonical correlation analysis (CCA) [65] is to project two sets of vectors into a joint subspace and then find a sequence of linear transformations of each set of variables that maximizes the relationship between the transformed variables. CCA values can be used to measure the similarity between the generated gestures and the real ones. The closer the CCA is to 1, the better. The Fréchet gesture distance (FGD) [88] on feature space is proposed as a metric to quantify the quality of the generated gestures. To compute the FGD, we trained an autoencoder to extract the feature. Lower FGD is better. Diversity [45] in feature space is used to evaluate the diversity of the gestures. We also report average jerk, average acceleration [35], Hellinger distance [36], and Beat Align Score [45, 50] in the supplementary material.

4.2 Comparison to Existing Methods

4.2.1 Objective Evaluation. We compare our proposed model with StyleGestures [4], Audio2Gestures [43], ExampleGestures [19], and DiffuseStyleGesture [85]. The quantitative results are shown in Table 1. On the global CCA, our proposed model outperforms all other existing methods. The highest global CCA shows a strong coupling between the generated gestures and the ground truth gestures. CCA for each sequence is not as good as the other methods, and we suggest that this is because for each speech, the model learns the gestures across the skeleton. Our method significantly surpasses the compared state-of-the-art methods with FGD, improves 6.64 (63%) than the best compared baseline model ExampleGestures. This shows the high quality of the generated gestures. We can see that our model is not as good as StyleGesture in terms of Diversity. The video results show that StyleGesture has a lot of cluttered movements, increasing diversity while decreasing human-likeness and appropriateness. However, we would like to emphasize that objective evaluation is currently not particularly relevant for assessing gesture generation [37]. Subjective evaluation remains the gold standard for comparing gesture generation models [37, 39]. Current research on speech-driven gestures prefers to conduct only subjective evaluation [5, 85]. Please refer to the supplementary video for more comparisons.

4.2.2 User Study. To understand the real visual performance of our method, we conduct a user study among the gesture sequences generated by each compared method and the ground truth motion capture data. Following the evaluation in GENEA [52], we evaluate human-likeness and gesture-speech appropriateness. The length of the evaluated clips ranged from 22 to 50 seconds, with an average length of 35.4 seconds, as longer durations produce more pronounced and convincing appropriateness results [87]. For human-likeness evaluation, each evaluation page asked participants "How human-like does the gesture motion appear?" In terms of appropriateness evaluation, each evaluation page asked participants "How appropriate are the gestures for the speech?" Participants rated at 1-point interval from 5 to 1, with labels (from best to worst) of "excellent", "good", "fair", "poor", and "bad". More details about the user study are shown in the supplementary material. The mean opinion scores (MOS) on human-likeness and appropriateness are reported in the last two columns in Table 1.

In terms of human-likeness, our model significantly surpasses the compared state-of-the-art methods. However, it is not significantly different from ExampleGestures. This is because ExampleGestures uses a reference gesture as 'example' during inference, with already a priori knowledge of the gesture, sampling from a gesture distribution to get the generated gesture, so the human-likeness is strong. For gesture and speech appropriateness, our model significantly outperforms StyleGestures, Audio2Gesture, and ExampleGestures, giving competitive results with DiffuseStyleGesture. One reason for

Table 1: Quantitative results on test set. Bold indicates the best metric. Among compared methods, StyleGestures [4], Audio2Gestures [43], ExampleGestures [19], and DiffuseStyleGesture [85] are reproduced using officially released code with some optimized settings. Objective evaluation is recomputed using the officially updated evaluation code [37, 45]. Human-likeness and appropriateness are the results of MOS with 95% confidence intervals.

Name	Objective evaluation				Subjective evaluation	
	Global CCA	CCA for each sequence	FGD ↓	Diversity ↑	Human-likeness	Appropriateness
Ground Truth	1.000	1.00 ± 0.00	0.0	10.03	4.22 ± 0.11	4.22 ± 0.11
StyleGestures [4]	0.978	**0.98 ± 0.01**	15.89	**13.86**	3.56 ± 0.12	3.17 ± 0.13
Audio2Gesture [43]	0.969	0.97 ± 0.01	19.78	6.148	3.61 ± 0.11	3.15 ± 0.14
ExampleGestures [19]	0.914	**0.98 ± 0.01**	10.49	5.418	3.77 ± 0.12	3.17 ± 0.14
DiffuseStyleGesture [85]	0.987	0.97 ± 0.01	11.98	11.22	3.66 ± 0.12	**3.46 ± 0.14**
Ours	**0.988**	0.95 ± 0.02	**3.850**	7.039	**3.80 ± 0.11**	3.42 ± 0.14

the gap compared to DiffuseStyleGesture is that DiffuseStyleGesture uses kinematic parameters such as the position, rotation angle, velocity, and rotation angular velocity of the root, as well as the position, rotation angle, velocity, rotation angular velocity, and gaze direction of each joint of the original motion as features of the gesture, which has a much larger dimension than the feature dimension of the primal skeleton gesture and may contain fine-grained skeletal details related to speech. According to the feedback from the participants, our generated gestures are "more semantically relevant" and "more natural", while our method has "less power" compared to Ground Truth. We suggest that this observation is due to the downsampling in the retargeting network and the VQVAE network. Smaller downsampling coefficients may result in faster and more powerful movements.

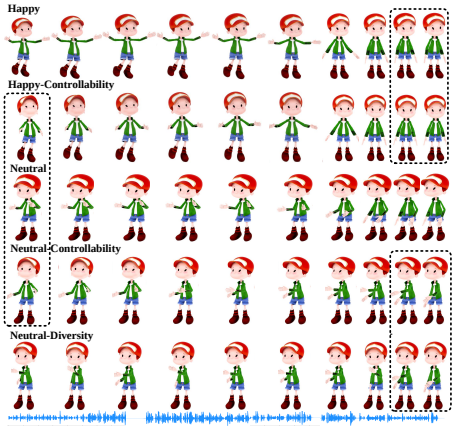

Figure 7: Visualization of the stylization, controllability, and diversity of generated gestures. We randomly select a 2.67-second generated gesture clip (10 codes). Then setting γ and s in Equation (13) to control the style and setting noisy gesture in diffusion model to generate diverse gestures. The dashed boxes indicate that we control their code the same.

4.3 Ablation Studies

4.3.1 Objective Evaluation. Moreover, we conduct ablation studies to address the performance effects of different components in the framework. We performed the experiments on the following components: (1) reinforcement learning, (2) VQVAE, and (3) multiple skeletons. The results of our ablation studies are summarized in Table 2. The metrics on FGD indicate that after RL finetuning, the generated gestures have increased distance from the distribution of human gestures in the dataset, indicating that the model has explored some gestures that do not belong to the existing distribution of gestures in the dataset but are considered reasonable by the reward model. From the CCA and diversity metrics, it can be seen that the reward model can indeed generalize to gestures outside the dataset, allowing the model to generate more diverse and high-quality gesture movements that are not limited to the dataset. When neither RL nor VQVAE is used, both FGD and diversity are still decreasing, which indicates the necessity of codebooks to generalize meaningful gestures. When we use only a single dataset, we notice that both FGD and diversity decrease a lot, which indicates the essential importance of gesture generation for learning on multiple datasets.

4.3.2 User Study. Similarly, we conduct a user study of ablation studies. The MOS on human-likeness and appropriateness are shown in the last two columns in Table 2. In terms of human similarity, we can find that the scale of the dataset has a significant effect on the results, which demonstrates the importance of unifying the gesture dataset. For speech and gesture appropriateness, it is also found that the scale of the dataset has the largest impact on this metric. Secondly, the appropriateness also decreased without reinforcement learning, shows the importance of data exploration. The visual comparisons of this study can be also referred to the supplementary video.

4.4 Diverse, Controllable, and Stylized Gesture Generation

- **Stylization.** We can generate stylized gestures by setting γ and s in Equation (13). The intensity of the stylization can be controlled by the value of γ. As shown in Figure 7, for the same speech, different styles of gestures can be generated while preserving matching with the speech.
- **Diversity.** Due to the diffusion model architecture, different noisy gesture and different seed gesture could generate different gestures even for the same speech and style, as

Table 2: Ablation studies results. '–' indicates modules that are not used. Bold indicates the best metric.

Name	Objective evaluation				Subjective evaluation	
	Global CCA	CCA for each sequence	FGD ↓	Diversity ↑	Human-likeness	Appropriateness
Ground Truth	1.000	1.00 ± 0.00	0.0	10.03	4.22 ± 0.11	4.22 ± 0.11
Ours	**0.988**	**0.95 ± 0.02**	3.850	**7.039**	3.80 ± 0.11	**3.42 ± 0.14**
- RL	0.987	0.94 ± 0.03	**3.132**	7.008	**3.82 ± 0.11**	3.24 ± 0.16
- RL - VQVAE	0.987	0.94 ± 0.03	3.568	6.971	3.79 ± 0.11	3.33 ± 0.12
- Skeleton A	0.972	0.94 ± 0.03	13.76	4.882	3.54 ± 0.12	3.00 ± 0.13
- Skeleton B	0.965	**0.95 ± 0.03**	12.45	5.566	3.59 ± 0.13	3.09 ± 0.13

shown in Figure 7. This is the same as real human speech, which creates diverse co-speech gestures related to the initial position.

- **Controllability.** Since we use VQVAE to generate gestures, it is easy to control the gesture or take out the code for interpretation. We can have a high level of control over speech-driven gestures at any time with the specified upper body code, as shown in the dashed box in 7.

For more details please refer to the supplementary material.

5 DISCUSSION AND CONCLUSION

In this paper, we assume that the body gestures of the different skeletons are contained in the primal skeleton and present a unified gesture synthesis model for multiple skeletons. UnifiedGesture demonstrates x major strength: 1) Benefit from using the skeleton-aware retargeting network to unify the different skeletons, while extending the dataset. The model has stronger generalization. And ablation experiments on a single skeleton effectively demonstrate that a larger amount of data can improve the performance of the model. 2) Based on a diffusion model, probabilistic mapping enhances diversity while enabling the generation of high-quality, speech-matched, and style-controlled gestures. 3) VQVAE learns a codebook to summarize meaningful gesture units to improve controllability and interpretability. Reinforcement learning with a learned reward function helps refine the gesture generation model, enabling the model to explore the data and able to increase the diversity of the generated gestures. The physics-based kinematic constraints also further improve gesture generation. There is room for improvement in this research. Besides speech, more modalities (e.g. text, facial expressions) could be taken into consideration to generate more appropriate gestures. Solving the problem that the skeleton-aware encoder and decoder need to be re-trained for the new skeleton is also our future research direction.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001, JCYJ20220818101014030) and Shenzhen Key Laboratory of next generation interactive media innovative technology (ZDSYS20210623092001004).

REFERENCES

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020. Skeleton-aware networks for deep motion retargeting. *ACM Trans. Graph.* 39, 4 (2020), 62. https://doi.org/10.1145/3386569.3392462

[2] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*. Association for Computational Linguistics, 1884–1895. https://doi.org/10.18653/v1/2020.findings-emnlp.170

[3] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. 2020. Style Transfer for Co-speech Gesture Animation: A Multi-speaker Conditional-Mixture Approach. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII (Lecture Notes in Computer Science, Vol. 12363)*. Springer, 248–265. https://doi.org/10.1007/978-3-030-58523-5_15

[4] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. *Comput. Graph. Forum* 39, 2 (2020), 487–496. https://doi.org/10.1111/cgf.13946

[5] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2022. Listen, denoise, action! Audio-driven motion synthesis with diffusion models. *CoRR* abs/2211.09707 (2022). https://doi.org/10.48550/arXiv.2211.09707 arXiv:2211.09707

[6] Simon Alexanderson, Éva Székely, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. 2020. Generating coherent spontaneous speech and gesture from text. In *IVA '20: ACM International Conference on Intelligent Virtual Agents, Virtual Event, Scotland, UK, October 20-22, 2020*. ACM, 1:1–1:3. https://doi.org/10.1145/3383652.3423874

[7] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic Gesticulator: Rhythm-Aware Co-Speech Gesture Synthesis with Hierarchical Neural Embeddings. *ACM Trans. Graph.* 41, 6 (2022), 209:1–209:19. https://doi.org/10.1145/3550454.3555435

[8] Autodesk. 2023. Maya. https://www.autodesk.com.cn/products/maya/overview.

[9] Andrew G Barto, Richard S Sutton, and Charles W Anderson. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics* 5 (1983), 834–846.

[10] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. AudioLM: a Language Modeling Approach to Audio Generation. *CoRR* abs/2209.03143 (2022). https://doi.org/10.48550/arXiv.2209.03143 arXiv:2209.03143

[11] Daniel S Brown, Wonjoon Goo, and Scott Niekum. 2020. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*. PMLR, 330–359.

[12] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* 16, 6 (2022), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

[13] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. 2018. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[14] Gabriel Dulac-Arnold, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, and Ben Coppin. 2015. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679* (2015).

[15] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X Liu. 2023. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature* 615, 7953 (2023), 620–627.

[16] Ylva Ferstl and Rachel McDonnell. 2018. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents, IVA 2018, Sydney, NSW, Australia, November 05-08, 2018*. ACM, 93–98. https://doi.org/10.1145/3267851.3267898

[17] Ylva Ferstl, Michael Neff, and Rachel McDonnell. 2019. Multi-objective adversarial gesture generation. In *Motion, Interaction and Games, MIG 2019, Newcastle upon*

Tyne, UK, October 28-30, 2019. ACM, 3:1–3:10. https://doi.org/10.1145/3359566.3360053

[18] Blender Foundation. 2023. Blender. https://www.blender.org/.

[19] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F. Troje, and Marc-André Carbonneau. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. *Comput. Graph. Forum* 42, 1 (2023), 206–216. https://doi.org/10.1111/cgf.14734

[20] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 3497–3506. https://doi.org/10.1109/CVPR.2019.00361

[21] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV (Lecture Notes in Computer Science, Vol. 13695)*. Springer, 580–597. https://doi.org/10.1007/978-3-031-19833-5_34

[22] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. 2018. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905* (2018).

[23] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. 2022. A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference*. 46:1–46:9. https://doi.org/10.1145/3528233.3530750

[24] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. 2021. Learning Speech-driven 3D Conversational Gestures from Video. In *IVA '21: ACM International Conference on Intelligent Virtual Agents, Virtual Event, Japan, September 14-17, 2021*. ACM, 101–108. https://doi.org/10.1145/3472306.3478335

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

[26] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *CoRR* abs/2207.12598 (2022). https://doi.org/10.48550/arXiv.2207.12598 arXiv:2207.12598

[27] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. *ACM Trans. Graph.* 41, 4 (2022), 161:1–161:19. https://doi.org/10.1145/3528223.3530094

[28] Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics*. 492–518.

[29] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2022. FLAME: Free-form Language-based Motion Synthesis & Editing. *CoRR* abs/2209.00349 (2022). https://doi.org/10.48550/arXiv.2209.00349 arXiv:2209.00349

[30] SangBin Kim, Inbum Park, Seongsu Kwon, and JungHyun Han. 2020. Motion Retargetting based on Dilated Convolutions and Skeleton-specific Loss Functions. *Comput. Graph. Forum* 39, 2 (2020), 497–507. https://doi.org/10.1111/cgf.13947

[31] Seong Uk Kim, Hanyoung Jang, and Jongmin Kim. 2020. A variational U-Net for motion retargeting. *Comput. Animat. Virtual Worlds* 31, 4-5 (2020). https://doi.org/10.1002/cav.1947

[32] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR*. http://arxiv.org/abs/1412.6980

[33] Michael Kipp. 2003. *Gesture generation by imitation: from human behavior to computer character animation*. Ph. D. Dissertation. Saarland University, Saarbrücken, Germany. http://scidok.sulb.uni-saarland.de/volltexte/2007/1256/

[34] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. In *8th International Conference on Learning Representations, ICLR April 26-30*. https://openreview.net/forum?id=rkgNKkHtvB

[35] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents, IVA 2019, Paris, France, July 2-5, 2019*. ACM, 97–104. https://doi.org/10.1145/3308532.3329472

[36] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *ICMI: International Conference on Multimodal Interaction*. 242–250. https://doi.org/10.1145/3382507.3418815

[37] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENEA Challenge 2020. In *26th International Conference on Intelligent User Interfaces*. 11–21. https://doi.org/10.1145/3397481.3450692

[38] Taras Kucherenko, Rajmund Nagy, Michael Neff, Hedvig Kjellström, and Gustav Eje Henter. 2022. Multimodal Analysis of the Predictability of Hand-gesture Properties. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 770–779. https://doi.org/10.5555/3535850.3535937

[39] Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. Evaluating gesture-generation in a large-scale open challenge: The GENEA Challenge 2022. *CoRR* abs/2303.08737 (2023). https://doi.org/10.48550/arXiv.2303.08737 arXiv:2303.08737

[40] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*.

[41] Gilwoo Lee, Zhiwei Deng, Shugao Ma, Takaaki Shiratori, Siddhartha S. Srinivasa, and Yaser Sheikh. 2019. Talking With Hands 16.2M: A Large-Scale Dataset of Synchronized Body-Finger Motion and Audio for Conversational Motion Analysis and Synthesis. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 763–772. https://doi.org/10.1109/ICCV.2019.00085

[42] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. 2020. Learning quadrupedal locomotion over challenging terrain. *Science robotics* 5, 47 (2020), eabc5986.

[43] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 11273–11282. https://doi.org/10.1109/ICCV48922.2021.01110

[44] Shujie Li, Lei Wang, Wei Jia, Yang Zhao, and Liping Zheng. 2022. An iterative solution for improving the generalization ability of unsupervised skeleton motion retargeting. *Comput. Graph.* 104 (2022), 129–139. https://doi.org/10.1016/j.cag.2022.04.001

[45] Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 11040–11049. https://doi.org/10.1109/CVPR52688.2022.01077

[46] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. 2022. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. arXiv:2208.00571 [cs.CV]

[47] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. 2022. SEEG: Semantic Energized Co-speech Gesture Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10463–10472. https://doi.org/10.1109/CVPR52688.2022.01022

[48] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. 2019. PMnet: Learning of Disentangled Pose and Movement for Unsupervised Motion Retargeting. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 136. https://bmvc2019.org/wp-content/uploads/papers/0997-paper.pdf

[49] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-modal Dataset for Conversational Gestures Synthesis. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 13667)*. Springer, 612–630. https://doi.org/10.1007/978-3-031-20071-7_36

[50] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. 2022. Learning Hierarchical Cross-Modal Association for Co-Speech Gesture Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10452–10462. https://doi.org/10.1109/CVPR52688.2022.01021

[51] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR, May 6-9*. https://openreview.net/forum?id=Bkg6RiCqY7

[52] Shuhong Lu and Andrew Feng. 2022. The DeepMotion entry to the GENEA Challenge 2022. In *International Conference on Multimodal Interaction, ICMI 2022, Bengaluru, India, November 7-11, 2022*. ACM, 790–796. https://doi.org/10.1145/3536221.3558059

[53] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. 2022. PoseGPT: Quantization-Based 3D Human Motion Generation and Forecasting. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 13666)*. Springer, 417–435. https://doi.org/10.1007/978-3-031-20068-7_24

[54] Jianxin Ma, Shuai Bai, and Chang Zhou. 2022. Pretrained Diffusion Models for Unified Human Motion Synthesis. *CoRR* abs/2212.02837 (2022). https://doi.org/10.48550/arXiv.2212.02837 arXiv:2212.02837

[55] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[56] Simbarashe Nyatsanga, Taras Kucherenko, Chaitanya Ahuja, Gustav Eje Henter, and Michael Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *CoRR* abs/2301.05339 (2023). https://doi.org/10.48550/arXiv.2301.05339 arXiv:2301.05339

[57] OpenAI. 2023. GPT-4 Technical Report. *CoRR* abs/2303.08774 (2023). https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774

[58] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[59] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. 2020. Modeling Human Motion with Quaternion-Based Neural Networks. *Int. J. Comput. Vis.* 128, 4 (2020), 855–872. https://doi.org/10.1007/s11263-019-01245-6

[60] André Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai. 2023. Tuning computer vision models with task rewards. *arXiv preprint arXiv:2302.08242* (2023).

[61] Xingqun Qi, Chen Liu, Muyi Sun, et al. 2023. Diverse 3D Hand Gesture Prediction from Body Dynamics by Bilateral Hand Disentanglement. *CoRR* abs/2303.01765 (2023). arXiv:2303.01765

[62] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. 2021. Speech Drives Templates: Co-Speech Gesture Synthesis with Learned Templates. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 11057–11066. https://doi.org/10.1109/ICCV48922.2021.01089

[63] Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. 2022. Diffusion Motion: Generate Text-Guided 3D Human Motion by Diffusion Model. *CoRR* abs/2210.12315 (2022). https://doi.org/10.48550/arXiv.2210.12315 arXiv:2210.12315

[64] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient Content-Based Sparse Attention with Routing Transformers. *Trans. Assoc. Comput. Linguistics* 9 (2021), 53–68. https://doi.org/10.1162/tacl_a_00353

[65] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Commun.* 110 (2019), 90–100. https://doi.org/10.1016/j.specom.2019.04.005

[66] Sebastian Starke, Ian Mason, and Taku Komura. 2022. DeepPhase: periodic autoencoders for learning motion phase manifolds. *ACM Trans. Graph.* 41, 4 (2022), 136:1–136:13. https://doi.org/10.1145/3528223.3530178

[67] Mingyang Sun, Mengchen Zhao, Yaqing Hou, Minglei Li, Huang Xu, Songcen Xu, and Jianye Hao. 2023. Co-speech Gesture Synthesis by Reinforcement Learning with Contrastive Pre-trained Rewards. (2023).

[68] Richard S Sutton. 1995. Generalization in reinforcement learning: Successful examples using sparse coarse coding. *Advances in neural information processing systems* 8 (1995).

[69] Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement learning - an introduction*. MIT Press. https://www.worldcat.org/oclc/37293240

[70] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* 12 (1999).

[71] Yunhao Tang and Shipra Agrawal. 2020. Discretizing Continuous Action Space for On-Policy Optimization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 5981–5988. https://ojs.aaai.org/index.php/AAAI/article/view/6059

[72] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. *CoRR* abs/2209.14916 (2022). https://doi.org/10.48550/arXiv.2209.14916 arXiv:2209.14916

[73] Jonathan Tseng, Rodrigo Castellon, and C. Karen Liu. 2022. EDGE: Editable Dance Generation From Music. *CoRR* abs/2211.10658 (2022). https://doi.org/10.48550/arXiv.2211.10658 arXiv:2211.10658

[74] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 6306–6315. https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[76] Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. 2021. Contact-Aware Retargeting of Skinned Motion. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 9700–9709. https://doi.org/10.1109/ICCV48922.2021.00958

[77] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. 2018. Neural Kinematic Networks for Unsupervised Motion Retargetting. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 8639–8648. https://doi.org/10.1109/CVPR.2018.00901

[78] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[79] He Wang, Edmond S. L. Ho, Hubert P. H. Shum, and Zhanxing Zhu. 2021. Spatio-Temporal Manifold Learning for Human Motions via Long-Horizon Modeling. *IEEE Trans. Vis. Comput. Graph.* 27, 1 (2021), 216–227. https://doi.org/10.1109/TVCG.2019.2936810

[80] Siyang Wang, Simon Alexanderson, Joakim Gustafson, Jonas Beskow, Gustav Eje Henter, and Éva Székely. 2021. Integrated Speech and Gesture Synthesis. In *ICMI '21: International Conference on Multimodal Interaction, Montréal, QC, Canada, October 18-22, 2021*, Zakia Hammal, Carlos Busso, Catherine Pelachaud, Sharon L. Oviatt, Albert Ali Salah, and Guoying Zhao (Eds.). ACM, 177–185. https://doi.org/10.1145/3462244.3479914

[81] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. ACM, 2022–2032. https://doi.org/10.1145/3308558.3313562

[82] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8 (1992), 279–292.

[83] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning* (1992), 5–32.

[84] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. 2022. Vector Quantized Diffusion Model with CodeUnet for Text-to-Sign Pose Sequences Generation. *CoRR* abs/2208.09141 (2022). https://doi.org/10.48550/arXiv.2208.09141 arXiv:2208.09141

[85] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023, Macao, S.A.R, 19th-25th August 2023*. ijcai.org.

[86] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver Canada, June 18-22, 2023*. IEEE.

[87] Sicheng Yang, Zhiyong Wu, Minglei Li, Mengchen Zhao, Jiuxin Lin, Liyang Chen, and Weihong Bao. 2022. The ReprGesture entry to the GENEA Challenge 2022. In *International Conference on Multimodal Interaction, ICMI, November 7-11*. 758–763. https://doi.org/10.1145/3536221.3558066

[88] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Trans. Graph.* 39, 6 (2020), 222:1–222:16. https://doi.org/10.1145/3414685.3417838

[89] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*. IEEE, 4303–4309. https://doi.org/10.1109/ICRA.2019.8793720

[90] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *International Conference on Multimodal Interaction, ICMI 2022, Bengaluru, India, November 7-11, 2022*. ACM, 736–747. https://doi.org/10.1145/3536221.3558058

[91] Fan Zhang, Naye Ji, Fuxing Gao, and Yongping Li. 2023. DiffMotion: Speech-Driven Gesture Synthesis Using Denoising Diffusion Model. In *MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13833)*. Springer, 231–242. https://doi.org/10.1007/978-3-031-27077-2_18

[92] Chi Zhou, Tengyue Bian, and Kang Chen. 2022. GestureMaster: Graph-based Speech-driven Gesture Generation. In *International Conference on Multimodal Interaction, ICMI*. 764–770. https://doi.org/10.1145/3536221.3558063

[93] Zixiang Zhou and Baoyuan Wang. 2022. UDE: A Unified Driving Engine for Human Motion Generation. *CoRR* abs/2211.16016 (2022). https://doi.org/10.48550/arXiv.2211.16016 arXiv:2211.16016

[94] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. *CoRR* abs/2303.09119 (2023). https://doi.org/10.48550/arXiv.2303.09119 arXiv:2303.09119

[95] Wentao Zhu, Zhuoqian Yang, Ziang Di, Wayne Wu, Yizhou Wang, and Chen Change Loy. 2022. MoCaNet: Motion Retargeting In-the-Wild via Canonicalization Networks. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 3617–3625. https://ojs.aaai.org/index.php/AAAI/article/view/20274

A OVERVIEW

In this supplementary file, we present more experimental results analysis.

- We illustrate the differences between the skeletons of different gesture datasets.
- We give the specific structure of the retargeting network.
- We describe the implementation of reinforcement learning in detail.
- We give numerical results for more objective metrics.
- We show the design and scoring interface of the user study.
- We illustrate the model inference process.

B SKELETONS OF DIFFERENT GESTURE DATASETS

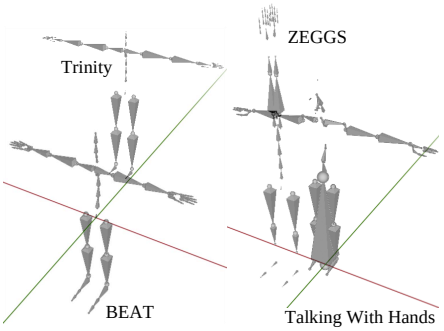

Figure 8: Visualization of different gesture datasets with reference poses, the position of their roots, and the number and location of their joints are different. The red and green lines represent the x-axis and y-axis, respectively.

In practice, the dataset of the target skeleton is often small due to the expensive cost of motion capture, and we address the problem of how to utilize the existing datasets to improve the generation quality of the target skeleton. To this end, we design our algorithm to extend the dataset by unifying different skeletal datasets. We have explored the validity and necessity of unifying different skeletal datasets (Trinity and Zeggs), which is our main contribution, and our attempt takes an initial step towards generating gestures with big data and large motion models in the future.

Our current experimental setup is mainly to verify the feasibility of our idea, and we chose not to include larger datasets (e.g., Talking With Hands and BEAT) in the prior experiments due to the following two reasons: 1) As shown in Figure 8, BEAT and Trinity have the same skeletal standard (Vicon), but Trinity, Zeggs, and Talking With Hands do not have the same skeleton. We experiment on Trinity and Zeggs, which are of comparable size, for the balance of the dataset distribution. 2) BEAT [49] has demonstrated better performance and generalization based on larger data compared to Trinity. That is, the more data with the same skeleton, the better the results. We solved the issue of adding data to improve performance from a different perspective, i.e., by adding more data with different skeletons.

The finger mocap is still very challenging currently in the industry: 1) Optical motion capture systems, such as Vicon and OptiTrack, have occlusion problems, the markers on hand are easy to be occluded by hands; 2) Inertial motion capture systems, such as Xsens and Noitom, have the error accumulating problems, the inertial sensors are easy to accumulate errors; And 3) some deep learning-based pose estimation algorithms use optical motion capture or inertial motion capture as Ground Truth, which causes even lower accuracy.

We found the hand/finger quality of the existing mocap datasets [56] is not good enough, especially when retargeted to an avatar. Datasets claimed with high-quality hand motion capture were still reported to have poor hand motion [56], e.g., ZEGGS Dataset in [5] and Talking With Hands in [90]. So we ignore hand/finger motion currently, and leave it to future work.

The skeleton of the different gesture datasets is shown in Figure 8. These reference poses are, for the Trinity dataset, the root is not at the xy-plane origin; for the BEAT dataset, the root is at the origin, not the feet at the origin; for the ZEGGS dataset, the reference pose is not T-pose, in a straight line, and also not at the origin; for the Talking With Hands dataset, the skeleton structure is more complex, with many small skeletons and relying on a 'World' skeleton to maintain body height. Like humans, in order to tell models what the references for retargeting are, we need to set up uniform reference gestures for them.

C RETARGETING NETWORK

In this paper, we aim to solve the issue of co-speech generation, not retargeting. Ablation experiments of skeleton-aware networks (e.g. simple CNN) and comparisons with other baseline models are beyond the scope of our paper and have been done in [1] to demonstrate the effectiveness of skeleton-aware networks for retargeting. Moreover, refinement after generating results is a general strategy, such as Bailando [45], GPT4 [57], and root and feet processing [1, 92].

We consider the retargeting module, diffusion generation module, and refinement module in this work as an integrated pipeline, and we have demonstrated the effectiveness of each module in ablation experiments, and the refinement module based on reinforcement learning and physics guidance can help to achieve better performance.

The network structure of the dynamic encoder $E^{dynamic}$ is visualized in Figure 9. It replicates the static offsets and concatenates them with dynamic motion, following skeletal convolution, LeakyRELU, and skeletal pooling. The network structure of the static encoder E^{static} and decoder D is similar to it. The decoder replaces the skeletal pooling with skeletal upsampling.

D HYPER-PARAMETERS OF REINFORCEMENT LEARNING

The hyper-parameter settings during our reinforcement learning process are shown in Table 3.

E OBJECTIVE EVALUATION

Regarding the models DisCo [101] and Speech2AffectiveGestures [97], only the upper body gestures were generated in their original papers. To generate gestures for the full body, these models need to learn kinematically relevant information, otherwise root sliding

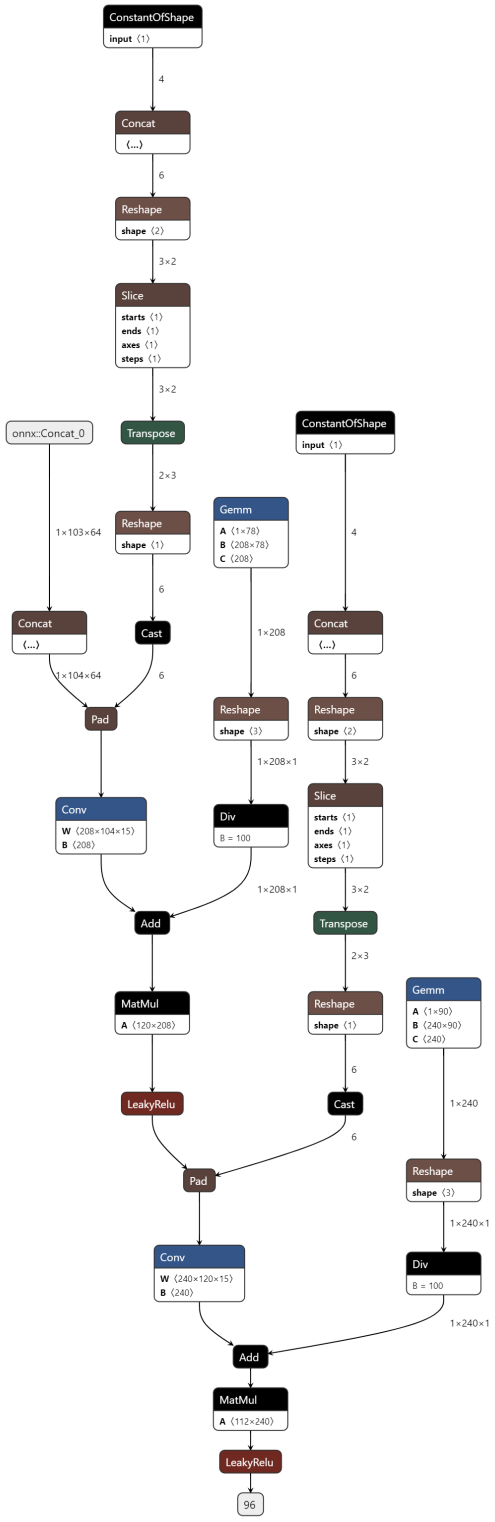

Figure 9: Visualization of the dynamic encoder $E^{dynamic}$ network structure. It mainly includes skeletal convolution, activation function, and skeletal pooling.

Table 3: Hyper-parameters of the reinforcement learning.

	Hyper-parameter	Value
Reward Model	Gesture Emb Dim	512
	Music Emb Dim	1024
	Hidden Dim	768
	Block Num	1
	Head Num	12
	Context Length	18
	Causal	False
	Optimizer	Adam
	Learning Rate	5e-4
	Weight Decay	2.5e-3
	Batch Size	128
	K List	[0, 4, 9, 13, 18]
REINFORCE	gamma	1.0
	Max Grad Norm	0.1
	Optimizer	AdamW
	Learning Rate	1e-6

and drifting will occur. Text2gestures [98] is similar, its dataset is composed of the gestures of a seated, root-immobile human, and the ability to generate full-body free-motion gestures needs further validation. As for Mofusion [99] and GestureDiffuCLIP [96], they do not have open source code and are recently released work, so this work is not compared with them. The baseline models chosen in this paper are recently published and well behaved in full-body gestures.

E.1 Comparison to Existing Methods

In Table 1, as for CCA, our model shows a slightly lower CCA for each sequence performance compared to some other methods, specifically StyleGestures and ExampleGestures. Our model achieves a CCA of 0.95, which is slightly lower than the 0.98 achieved by StyleGestures and ExampleGestures. The primary reason is that our model is designed to learn gestures across the entire skeleton, thereby emphasizing global coherence and alignment. This global approach may lead to compromises in capturing the detailed correlations at individual sequence levels. It is important to note that this trade-off is partly intentional since our aim was to excel in global CCA (0.988), which reflects the strong coupling between the generated gestures and the ground truth. With respect to diversity, our model scores 7.039, which is not as high as StyleGestures' 13.86. This is attributable to the fact that our model focuses on achieving human-likeness and appropriateness in gestures, which sometimes necessitates generating gestures that are more constrained and less varied. While StyleGestures focuses more on producing diverse gestures, it doesn't perform as well in the FGD (Frechet Gesture Distance) metric, indicating that its gestures may not be as high quality as those generated by our model.

The additional objective measures compared to the baseline model are shown in Table 4. Average Jerk, Average Acceleration, and Hellinger Distance are recomputed using [37]. As for Beat Align Score, we use the method in [43] to calculate the beats of audio

and follow [45] to calculate the beats and diversity of gestures. For Average Jerk and Average Acceleration, the closer to Ground Truth, the better. For Hellinger Distance, the smaller the better. Regarding Beat Align Score, the greater the better. From the results, it can be seen that the gestures generated by our model are closest to the real velocity and acceleration distributions. StyleGestures and DiffuseStyleGesture have motion velocity histogram distances that are more similar to the real gestures. This could be caused by the more hand movements of both of them, please refer to our supplementary video. DiffuseStyleGesture matches the beat of the speech better, which is consistent with the results of the human subjective evaluation. Here we also want to emphasize that currently there is a lack of valid objective metrics for gesture generation and that subjective evaluation is the most effective [37, 39]. Please refer to our video for further visualization and comparison.

E.2 Ablation Studies

In Table 2, we observe that when the RL (Reinforcement Learning) component is removed from our model (Ours - RL), the FGD (Frechet Gesture Distance) decreases from 3.850 to 3.132. This indicates that the gestures generated without RL are closer to the distribution of human gestures in the dataset. However, the slightly better FGD score does not necessarily represent better generalization. RL is essential for enabling the model to explore beyond the dataset and generate gestures that, though slightly further from the human distribution, are more diverse and considered reasonable by the reward model, as can be seen from the Global CCA and Diversity metrics. Therefore, while the ablated version without RL shows better FGD, the trade-off is in generalization and diversity. When neither RL nor VQVAE (Vector Quantized Variational AutoEncoder) is used (Ours - RL - VQVAE), the FGD is higher than Ours - RL, but still lower than our full model. The absence of VQVAE causes a reduction in diversity. This suggests that the VQVAE module helps to generate meaningful gestures. In this ablation, without the RL module, the model is not encouraged to explore beyond the dataset, and without VQVAE, it struggles to generalize meaningful gestures. The combination of RL and VQVAE in the full model ensures that meaningful gestures are generated, and the model is encouraged to explore beyond the dataset, which enhances the diversity and quality of the generated gestures. The ablation studies with - Skeleton A and - Skeleton B demonstrate the importance of having diverse training datasets. As we can see, removing either Skeleton A or Skeleton B increases the FGD dramatically to 13.76 and 12.45 respectively. This indicates that the model is not generalizing well without diverse training data. Similarly, Diversity is significantly reduced when training on a single dataset, highlighting the importance of training on multiple datasets for producing varied and high-quality gestures. Our full model, incorporating RL, VQVAE, and multiple datasets, achieves a balance across these aspects, as is evident in the objective and subjective evaluations.

Similarly, we calculated more objective measures of the ablation experiment. The results are shown in Table 5. As can be seen from the results, all four metrics even get better when the model does not use reinforcement learning as well as VQVAE; when the model removes reinforcement learning and VQVAE, the average velocity, average acceleration, and velocity distribution histogram achieves the best case. When the model was trained on a single dataset skeleton only, the average velocity and average acceleration decreased more significantly, but the other two metrics showed the opposite trend. These are contradictory to the subjective evaluation and other objective evaluation metrics, so we argue with the results of these objective metrics, but we still report these results. We believe that subjective evaluation is still the most convincing method for now, and that objective metrics are all still inconsistent with subjective human perception. Please refer to the video for further comparison.

F USER STUDY

Human-likeliness and Appropriateness for gesture scoring are the two dimensions that have been used in the gesture generation (GENEA) Challenge [37, 39, 90] and are currently the dominant metrics in gesture generation. Some work in user study has different focus on different topics, such as user evaluation diversity [43], consistency [102], stylization [85], appropriateness of interaction with the listener added in this year's gesture generation Challenge [100], etc. However, Human-likeliness and Appropriateness are dimensions that are used in almost all user studies of gesture generation methods, so we followed these two metrics. To analyze user studies using statistics, we used MOS with 95% confidence intervals to represent the results of each metric. If there is no overlap in the 95% confidence intervals of the ratings between the different models, then the difference is considered to be statistically significant.

We put all the generated primal skeleton gestures through the decoder of the ZEGGS dataset skeleton to generate the final gestures. The generated motion capture file (bvh) is rendered by Blender [18] and the camera stays still. Before starting the evaluation, we told participants, for each Video, two dimensions were evaluated: 1. naturalness (human-likeness), i.e., the quality of the generated motion, without considering speech. 2. suitability (appropriateness), i.e., the relationship between the generated gestures and the speech, considering the speech, e.g., whether the gestures match the audio rhythm, or the text semantics. We asked people to ignore the influence of the hands on the scoring and to focus only on the skeletal movements of the whole body. A total of 31 individuals took part in the subjective evaluation scoring, with 2 subjects between the ages of 40 and 50 and the rest between the ages of 20 and 30. About 85% of the participants were male and 15% were female. They were all good English speakers. For both the comparison with the baseline model and the ablation experiments, we selected 10 segments of audio to be scored with their generated gestures. Five segments are male voices from the Trinity dataset, and another 5 segments are female voices from the ZEGGS dataset. Ten models in total were scored. For each model, there were 10 segments (approximately 5 minutes in total) of audio with generated gestures. We paid each participant an hourly rate of approximately 10 USD, which is above the average salary level [39]. A screenshot of the subjective evaluation scoring screen is shown in Figure 10.

G GESTURE GENERATION FOR MULTIPLE SKELETONS

Take the general gesture generation of skeleton A as an example. $\hat{\mathbf{L}}_0^{\text{upper}}$ and $\mathbf{L}_0^{\text{lower}}$ are the upper body of primal gesture sequence after VQVAE reconstruction and the lower body of the diffusion

Table 4: Additional quantitative results on the test set. Bold indicates the best metric, e.g., closest to Ground Truth or minimal, etc.

Name	Average Jerk	Average Acceleration	Hellinger Distance ↓	Beat Align Score ↑
Ground Truth	745.42 ± 661.80	64.50 ± 46.93	0.0	0.191
StyleGestures [4]	7667.90 ± 3609.80	319.40 ± 96.62	**0.139**	0.156
Audio2Gesture [43]	8460.02 ± 578.60	306.10 ± 25.91	0.151	0.123
ExampleGestures [19]	90.82 ± 5.69	18.81 ± 0.76	0.149	0.157
DiffuseStyleGesture [88]	1958.61 ± 316.69	204.71 ± 29.64	**0.139**	**0.239**
Ours	**263.87 ± 70.95**	**28.66 ± 9.26**	0.141	0.166

Table 5: Additional ablation studies results. '–' indicates modules that are not used. Bold indicates the best metric.

Name	Average Jerk	Average Acceleration	Hellinger Distance ↓	Beat Align Score ↑
Ground Truth	745.42 ± 661.80	64.50 ± 46.93	0.0	0.191
Ours	263.87 ± 70.95	28.66 ± 9.26	0.141	0.166
- RL	270.63 ± 75.24	29.30 ± 9.56	0.135	0.173
- RL - VQVAE	**276.83 ± 76.98**	**29.62 ± 9.69**	**0.126**	0.172
- Skeleton A	259.49 ± 67.10	26.20 ± 8.69	0.133	**0.176**
- Skeleton B	232.15 ± 52.80	20.45 ± 6.64	0.155	0.162

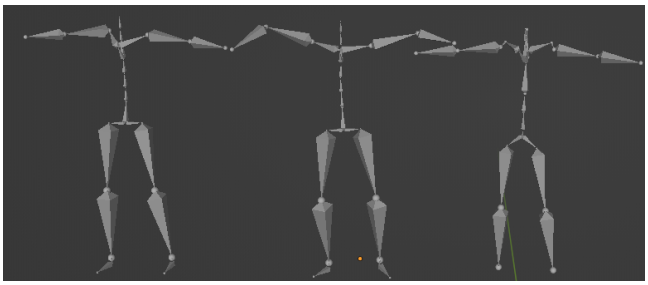

Figure 11: Retargeting visualization on BEAT and TWH. On the left is the result using Auto-rig in Blender, in the middle is the real motion, and on the right is the result generated by the retargeting network.

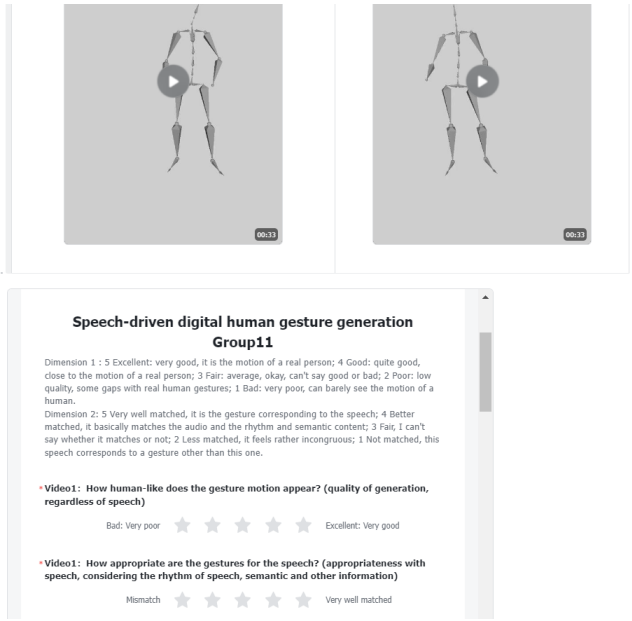

Figure 10: A screenshot of the subjective evaluation scoring interface.

model output, respectively. The generated gestures are finally given by Equation (4) as $D_A[(\hat{\mathbf{L}}_0^{\text{upper}}, \mathbf{L}_0^{\text{lower}}), \overrightarrow{\mathbf{S}_\mathbf{A}}]$. Similarly, the primal gesture sequence is fed into the decoder of whichever skeleton the gesture is generated, as shown in Figure 2.

H MORE DISCUSSION

We obtained similar results from our experiments on BEAT and TWH. The criteria for the motions in the BEAT and Trinity datasets are the same, and TWH is slightly more complex, especially for the shoulder joints, as shown on the right side of the Figure 11. From the figure, we can find that the three poses are generally similar, because the retargeting network is constrained with terminal positions and uses 5 terminals + 2 intermediate joints for a total of 7 joints as the middle representation, so more detailed information may be neglected. For example, the details of elbow and shoulder.

We appreciate the concerns regarding the scalability of our proposed method as new skeleton data is incorporated, necessitating the retraining of the network. We agree that it is theoretically feasible to decouple the problem into two separate parts - one focusing only on learning a uniform skeleton representation for the autoencoder system, and the other focusing only on co-speech gesture synthesis. This approach allows only the retargeting part of the network to be retrained when new skeleton data is added, resulting in significant computational cost and time savings. However, it is also important to acknowledge that while these two problems can be technically decoupled, there may be complex interactions between them in practice. There are indeed a lot of works [31, 35, 54, 77] on learning to retarget between different skeletons; or on learning co-speech gesture generation [6, 20, 50, 52]. We are the first approach

to attempt to integrate the both, and the integration of retargeting with speech-driven gestures can yield impressive results.

APPENDIX REFERENCES

[A96] Tenglong Ao, Zeyi Zhang, and Libin Liu. 2023. GestureDiffuCLIP: Gesture diffusion model with CLIP latents. *arXiv preprint arXiv:2303.14613* (2023).

[A97] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. 2021. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2027–2036.

[A98] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. 2021. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*. IEEE, 1–10.

[A99] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9760–9770.

[A100] Taras Kucherenko, Rajmund Nagy, Youngwoo Yoon, Jieyeon Woo, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. 2023. The GENEA Challenge 2023: A large scale evaluation of gesture generation models in monadic and dyadic settings. *arXiv preprint arXiv:2308.12646* (2023).

[A101] Haiyang Liu, Naoya Iwamoto, Zihao Zhu, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. DisCo: Disentangled Implicit Content and Rhythm Learning for Diverse Co-Speech Gestures Synthesis. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3764–3773.

[A102] Haolin Zhuang, Shun Lei, Long Xiao, Weiqin Li, Liyang Chen, Sicheng Yang, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2023. GTN-Bailando: Genre Consistent long-Term 3D Dance Generation Based on Pre-Trained Genre Token Network. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.